This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

NullSwap: Proactive Identity Cloaking Against Deepfake Face Swapping

Tianyi Wang[1] Shuaicheng Niu[1] Harry Cheng[2] Xiao Zhang[3] Yinglong Wang[3,*]
[1]Nanyang Technological University [2]Shandong University [3]Qilu University of Technology
{terry.ai.wang, xaCheng1996, zhang0510x}@gmail.com
shuaicheng.niu@ntu.edu.sg, wangyinglong@qlu.edu.cn

Abstract

Suffering from performance bottlenecks in passively detecting high-quality Deepfake images due to the advancement of generative models, proactive perturbations offer a promising approach to disabling Deepfake manipulations by inserting signals into benign images. However, existing proactive perturbation approaches remain unsatisfactory in several aspects: 1) visual degradation due to direct element-wise addition; 2) limited effectiveness against face swapping manipulation; 3) unavoidable reliance on white- and grey-box settings to involve generative models during training. In this study, we analyze the essence of Deepfake face swapping and argue the necessity of protecting source identities rather than target images, and we propose NullSwap, a novel proactive defense approach that cloaks source image identities and nullifies face swapping under a pure black-box scenario. We design an Identity Extraction module to obtain facial identity features from the source image, while a Perturbation Block is then devised to generate identity-guided perturbations accordingly. Meanwhile, a Feature Block extracts shallow-level image features, which are then fused with the perturbation in the Cloaking Block for image reconstruction. Furthermore, to ensure adaptability across different identity extractors in face swapping algorithms, we propose Dynamic Loss Weighting to adaptively balance identity losses. Experiments demonstrate the outstanding ability of our approach to fool various identity recognition models, outperforming state-of-the-art proactive perturbations in preventing face swapping models from generating images with correct source identities.

1. Introduction

The flourishing of deep generative models has brought significant development to Deepfake face manipulation techniques [18, 38]. Due to the improving visual quality of Deepfake images, the domain of passive Deepfake detection has reached bottlenecks in uncovering synthetic artifacts. Therefore, in recent years, researchers have moved on to proactively protect potential victim images in advance.

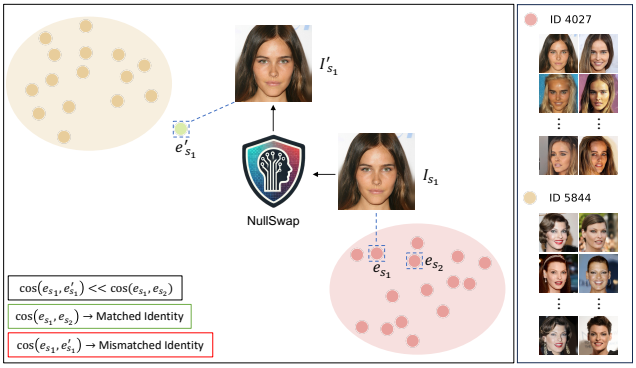


Figure 1. ArcFace [6] extracted embeddings of images for two facial identities, clustered via t-SNE [12]. NullSwap aims to insert the perturbation that blinds the identity extractors of face swapping models, such that the perturbed embedding (e'_{s_1}) no longer belongs to the original identity of embeddings (e_{s_1} and e_{s_2}).

While traditional passive detection approaches [3, 5, 8, 28, 30, 33, 34, 45] distinguish real and fake after Deepfake manipulations occur, proactive defense [16, 32, 35, 36, 40] against Deepfake aims to insert invisible signals into benign images in advance of potential Deepfake manipulations. Proactive perturbations, designed to be imperceptible to human eyes, act as countermoves to distort or nullify Deepfake face manipulations. The perturbations are usually trained to bring significant visual distortions and artifacts to synthetic results regarding certain types of Deepfake manipulations.

Besides reasonable defensive effects achieved by existing studies, several issues remain unresolved. First, although manipulations such as facial attribute editing and face reenactment have been promisingly handled [16, 27, 39], due to the additional difficulty, nullifying Deepfake face swapping remains largely unattempted and unachieved. Secondly, while most proactive algorithms aim to insert per-

*Corresponding author.

turbations into the target victim images, we argue the necessity to protect the source victim identities in face swapping, as the swapped identities are the ones in false scandals. Lastly, current solutions mostly rely on practical Deepfake models or imitated surrogate models when learning the perturbation patterns, which brings heavy computing burdens.

To resolve the aforementioned issues, we propose NullSwap, a novel identity cloaking algorithm designed to proactively defend against malicious Deepfake face swapping attacks. In general, when a celebrity or politician's face is swapped into compromising content, the individuals in the original videos may not require protection. Instead, the focus should be on safeguarding the identities whose faces are swapped, as they are the ones who ultimately face the repercussions of being embroiled in scandals and public controversies. Specifically, since most face swapping algorithms follow the pipeline of obtaining the identity information from the source image and fusing into the target image features for output image reconstruction and production, as demonstrated in Figure 1, we propose to invisibly cloak the underlying identity information such that incorrect identity embeddings are extracted and the face swapping results are then generated with undesired identities.

To proactively protect the source identities, for an input image, we first analyze its identity information via an Identity Extraction module. The extracted identity features are then fed to a Perturbation Block to derive the corresponding identity-guided perturbation for identity cloaking. On the other hand, we build a Feature Block to obtain image features from the input image at shallow-depths, pass them to a Cloaking Block along with the learned perturbation, and reconstruct the input image with the perturbation imperceptibly embedded. To ensure generalizability when encountering face swapping algorithms that adopt different identity extractors, we devise a Dynamic Loss Weighting (DLW) mechanism to adaptively adjust the total identity loss based on the values returned by the participating face recognition tools during training. Extensive experiments further demonstrate the ability of NullSwap to cloak the underlying identity information of the original source images and prevent malicious Deepfake face swapping from generating the desired identities even under a black-box scenario, such that no generative model participates during training. The contributions of this work are as follows:

- Upon analyzing the barely discussed essence of face swapping, we propose NullSwap, a novel idea to protect source images and prevent Deepfake face swapping from generating desired identities by embedding perturbations to cloak the identity information.
- To our knowledge, we are the first to proactively nullify Deepfake face swapping in a black-box scenario where generative models are omitted in training. Meanwhile, a novel Dynamic Loss Weighting strategy is devised to adaptively balance identity losses for generalizability.
- Experiments demonstrate the promising ability of NullSwap to cloak the underlying identity information against face recognition tools and to nullify Deepfake face swapping in deriving undesired identity.

2. Related Work

2.1. Deepfake Face Swapping

Deepfake face swapping performs identity transfer from a source image to the target one. Despite the original deepfakes[1] and several follow-up improvements [21, 25] that mainly utilize pure Auto-Encoder [20] in an identity-specific manner, the popular face swapping pipeline for arbitrary faces relies on obtaining the facial identity embeddings from the source faces and applying feature injection and fusion when editing the target faces. In 2020, a milestone work, SimSwap [4], extracts the identity embeddings as vectors from the source faces and introduces a novel identity injection module by adopting adaptive instance normalization (AdaIN) [17] to integrate features and accomplish target face reconstruction. Similarly, later approaches [9, 23, 41] follow the same pipeline while improving the visual performance with novel add-ons. For example, besides extracting identity embeddings via ArcFace [6], AP-Swap [37] provides an efficient global residual attribute-preserving encoder and a landmark-guided feature entanglement module to better preserve the identity-unrelated attributes.

2.2. Proactive Perturbation for Deepfake Detection

Proactive perturbations inserted into the original images are generally aimed to distort Deepfake generation. Most approaches [1, 31, 39, 42, 47] focus on protecting the victim image that is edited by Deepfake by direct element-wise addition of the perturbation to the image. Ruiz *et al*. [27] proposed a gradient-based method to add imperceptible perturbations to the image and have it distort the Deepfake results. Later, Huang *et al*. [16] raised the idea of Initiative defense by building a surrogate model to imitate Deepfake manipulation for devising a poison perturbation generator to obtain the venom for injection. To make it more compatible with human eyes, Wang *et al*. [32] explored Anti-Forgery to add perturbations to channels of the lab color space. Huang *et al*. [15] introduced the Cross-Model Universal Adversarial Watermark (CMUA), a cross-model universal attack pipeline, to attack multiple deepfake models iteratively, and fixed a perturbation generalized to all images and models. Dong *et al*. [7] proposed TCA-GAN to construct adversarial perturbation for disrupting unknown Deepfake systems via the assistance of a substitute model for face reconstruction.

[1] https://github.com/deepfakes/faceswap

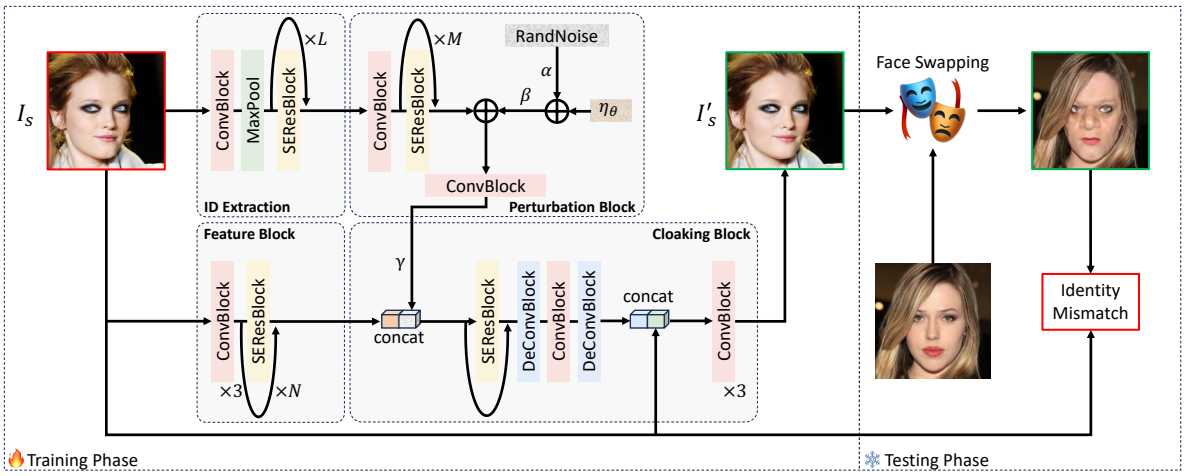


Figure 2. Demonstration of the proposed NullSwap framework. The input image I_s is passed through the ID Extraction module and Perturbation Block to generate identity-guided perturbation. A Feature Block executes shallow-level feature extraction on I_s and passes to the Cloaking Block together with the perturbation for the reconstruction of identity-cloaked $I_s^{'}$.

Qu *et al*. [26] designed DF-RAP to provide persistent protection for facial images under OSN compressions. A compression approximation GAN is trained to explicitly model OSN compression and then incorporated as a sub-module of the target Deepfake model.

However, existing literatures have barely considered the potential visual quality issue caused by direct element-wise addition [11, 44], and the popular white-box and grey-box settings have brought high computational costs during model training. Moreover, considering the actual victim of Deepfake face swapping, the protection of the source image as the identity provider has rarely been discussed.

3. Method

3.1. Problem Formulation

Ever since the first occurrence of Deepfake face swapping, various milestone studies [4, 28, 41] have advanced the research domain with better generative quality, but the general pipeline remains relatively fixed. In specific, for a source face that provides the desired facial identity and a target face whose facial identity is modified while maintaining the rest attributes, facial identity feature extraction usually happens to the former, and feature fusion followed by image reconstruction happens on the latter. In practice, a victim being face-swapped onto indelicate content is mired in scandal because of the high identity similarity as displayed in the synthetic media, somehow regardless of the visual quality. Therefore, although most existing proactive countermoves attempt to distort the generated facial images with different types of perturbations, we argue that, considering the essence of Deepfake face swapping, the goal of proactive defense is to prevent the face swapping models from extracting the correct identity information so that synthetic faces are generated with undesired and unexpected facial identities. In this way, Deepfake face swapping is proactively nullified while the generative models are untouched.

In this study, we propose NullSwap, a novel idea to achieve identity cloaking by embedding identity-guided perturbations into the source images, such that the perturbed images are visually identical to the original clean ones while the identity embeddings extracted by face recognition tools are vastly shifted (Figure 1). Since the objective can be achieved with the assistance of multiple face recognition tools, to the best of our knowledge, our proposed approach is the first proactive perturbation approach against face swapping in a black-box scenario without generative models participating during the training phase.

3.2. NullSwap Framework

For an unprotected source image I_s under the risk of malicious Deepfake face swapping, we propose NullSwap (Figure 2) to derive a perturbation protected image $I_s^{'}$. I_s is first passed to the Identity Extraction (ID Extraction) module for identity feature extraction, and then a Perturbation Block to derive the identity-guided perturbation that can cloak the identity information. Meanwhile, a Feature Block is applied to the input image I_s for shallow-level feature extraction. The image feature and perturbation are passed to a Cloaking Block and the reconstructed image $I_s^{'}$ is ultimately derived.

Identity Extraction. Unlike face recognition tools [6, 24, 29] that provide identity embeddings in vector format, we establish an ID Extraction module to analyze the input image and maintain identity features that are dominant toward

our objectives in matrix format. In specific, a ConvBlock, consisting of a CNN layer, a batch normalization, and a ReLU activation, followed by a max-pooling operation, is applied to I_s. To moderately maintain critical features from early stages and efficiently analyze the inter-channel correlation, we apply L consecutive SEResBlocks, where each block adopts the ResNet [10] bottleneck block and the squeeze-and-excitation networks (SENet) [13]. The feature sufficiency in this module provides the training capacity to consistently fool multiple face recognition tools.

Perturbation Block. The extracted identity features are fed to a Perturbation Block to derive identity-guided perturbation for the ultimate identity cloaking. Specifically, following a ConvBlock for feature refinement, M SEResBlocks are sequentially applied for hierarchical feature aggregation, preserving the identity features that are most relevant for perturbation and suppressing the less relevant ones. Meanwhile, considering the possible overfitting to specific identity patterns, we introduce a degree of randomness to the perturbation. In particular, after initializing a random noise drawn from the normal distribution, we derive the adaptive random noise following

$$\text{RandNoise} = \beta \cdot (\alpha \cdot \text{RandNoise} + \eta), \tag{1}$$

where η is a learnable noise, and α and β are learnable parameters to adaptively adjust the magnitude of randomness. The random noise is then added to the deterministic perturbation, as exhibited in Figure 2.

Feature Block. Considering the visual flaws caused by direct perturbation addition in existing work, we propose to embed the learned perturbation into latent image features. Therefore, for the input image I_s, we build a Feature Block to perform shallow-level feature extraction. Specifically, after three consecutive ConvBlocks for local feature analysis and dimension adjustment, we apply N SEResBlocks to enhance feature adaptability and context awareness before image reconstruction for $I_s^{'}$.

Cloaking Block. In this Cloaking Block, we reconstruct the input image while invisibly embedding the perturbation. First, by assigning a learnable weight γ to the perturbation, we concatenate it with the shallow-level image feature and apply a feature-level reconstruction via a SEResBlock, a DeConvBlock, a ConvBlock, and a DeConvBlock to balance perturbation embedding and visual fidelity. To elaborate, the DeConvBlock consists of the upsampling operation, batch normalization, and ReLU activation. In general, this feature-level reconstruction gradually restores spatial structures while effectively integrating the identity-guided perturbation.

In the end, for the feature-level reconstructed result containing a first-stage fusion of the perturbation and image features, we concatenate the input image to it and apply a final image-level reconstruction via three sequential ConvBlocks to derive $I_s^{'}$. With the two reconstructions at different dimension levels, we ensure that the reconstructed image remains perceptually similar to the original with mere artifacts while successfully cloaking the source identity.

3.3. Dynamic Loss Weighting

In this study, the main objective is to fool the face recognition tools upon identity embedding extraction. Since the adopted embedding extractor varies in different face swapping algorithms, we introduce multiple face recognition tools to ensure the generalizability of our approach. We design Dynamic Loss Weighting (DLW) to adaptively balance loss objectives drawn via different face recognition algorithms during training. It consists of two core components to determine the relative importance of each loss at each time state, namely, loss variance and relative progress. Loss variance measures the stability of each loss over a recent training window. Specifically, losses with higher variance, indicating instability or noise, are assigned lower weights to ensure stable optimization. On the other hand, relative progress evaluates the rate of improvement for each loss. In particular, losses showing slower progress are prioritized, allowing them to catch up and ensuring balanced contributions across tasks.

By introducing DLW, the identity loss is denoted as

$$\mathcal{L}_{\text{id}}(t_e, t_b) = \Sigma_{i=1}^{c} \hat{w}_i(t_e, t_b) \cdot \mathcal{L}_i(t_b), \tag{2}$$

where for each $\mathcal{L}_i$ of the total of c loss functions at iteration t_b of epoch t_e, the weight $\hat{w}_i$ is determined by

$$\hat{w}_i(t_e, t_b) = \frac{c \cdot w_i(t_e, t_b)}{\Sigma_{j=1}^{c} w_j(t_e, t_b)}, \tag{3}$$

which denotes the normalization process that aligns each weight w_i regarding the magnitude of different losses. Before normalization, the weight w_i is derived by

$$w_i(t_e, t_b) = \max\left(\frac{1}{\max(\alpha \cdot \sigma_i^2(t_b) + \beta(t_e) \cdot (1 + \Delta_i(t_b)), \epsilon_d)}, \epsilon_w\right), \tag{4}$$

where α is a constant regularization factor that balances the loss variance and relative progress. β relies on the current epoch t_e by

$$\beta(t_e) = \min\left(\beta_{\text{init}} + \gamma \cdot \min(t_e, t_e^*), \beta^*\right), \tag{5}$$

where t_e^* and β^* are the upper limit values for t_e and β, respectively. The variance σ_i^2 is computed based on the recent k consecutive loss values with respect to the current iteration t_b following

$$\sigma_i^2(t_b) = \frac{1}{k} \sum_{j=0}^{k-1} \left(\mathcal{L}_i(t_b - j) - \frac{1}{k} \Sigma_{m=0}^{k-1}(\mathcal{L}_i(t_b - m))\right)^2, \tag{6}$$

and Δ_i is the relative progress computed based on loss values of the current and the previous iterations

$$\Delta_i = \max\left(\frac{\mathcal{L}_i(t_b-1)-\mathcal{L}_i(t_b)}{\mathcal{L}_i(t_b-1)+\epsilon_p}, -1\right). \quad (7)$$

DLW ensures that losses contributing more stable and consistent gradients maintain their influence, while noisy or under-performing losses are appropriately adjusted. By combining loss variance and relative progress adaptively, DLW aligns the training process with the overarching goal of optimizing all objectives fairly. Furthermore, DLW is computationally efficient and directly operates on loss values, dynamically adjusting loss contributions without requiring additional parameters or gradient-level manipulation.

3.4. Objective Functions

In this study, the main goal is to maintain image reconstruction quality while promisingly nullifying face recognition tools. On the one hand, considering the pixel-wise similarity, we employ the Mean Squared Error (MSE) loss denoted as

$$\mathcal{L}_{\text{MSE}} = \|I_s - I_s^{'}\|_2. \quad (8)$$

On the other hand, regarding the potential artifacts that are negligible by MSE loss such as blurring and loss of high-frequency details, we adopt the Learned Perceptual Image Patch Similarity (LPIPS) loss [43] to measure perceptual similarity based on deep feature representations following

$$\mathcal{L}_{\text{LPIPS}} = \Sigma_l \|\phi_l(I_s) - \phi_l(I_s^{'})\|_2^2, \quad (9)$$

where ϕ_l represents the feature activation from a pre-trained AlexNet [22] at layer l.

Furthermore, we train a discriminator from scratch along with the NullSwap model training, where the discriminator acts as a binary classifier with N ConvBlocks. The objective function of the discriminator D is denoted as

$$\mathcal{L}_{\text{dis}} = -\mathbb{E}(\log D(I_s)) + \mathbb{E}\log(1 - D(I_s^{'})). \quad (10)$$

While the discriminator is iteratively trained, it provides guidance on the visual quality of perturbed images following

$$\mathcal{L}_D = -\mathbb{E}(\log D(I_s^{'})). \quad (11)$$

As for the performance in cloaking the identity information, to guarantee generalizability, we adopt multiple face recognition tools to provide the cosine similarities between identity embeddings e_s and $e_s^{'}$ extracted from I_s and $I_s^{'}$, respectively. The formula is denoted as

$$\mathcal{L}_i(e_s, e_s^{'}) = \frac{e_s \cdot e_s^{'}}{\|e_s\|\|e_s^{'}\|}, \quad (12)$$

Models	PSNR↑	SSIM↑	LPIPS↓
Initiative [16]	39.3803	0.9544	0.0200
Anti-Forgery [32]	38.0705	0.9530	0.0282
CMUA [15]	38.6395	0.9504	0.0333
DF-RAP [26]	38.8466	0.9349	0.0511
NullSwap (Ours)	**41.3097**	**0.9864**	**0.0049**

Table 1. Visual quality evaluation of perturbed images. Best performance marked in **bold**.

and the identity loss is computed following DLW as demonstrated in Eqn.(2).

The total loss is computed as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_D\mathcal{L}_D, \quad (13)$$

with pre-defined coefficients λ_{id}, λ_{MSE}, λ_{LPIPS}, and λ_D.

4. Experiments

4.1. Implementation Details

In the experiment, we adopted the popular human face dataset, CelebA-HQ [19], which contains 30,000 high-quality facial images with 6,217 unique identities, following the official split for training, validation, and testing. Meanwhile, we adopted another classical face dataset, Labeled Faces in the Wild (LFW) [14] with 5,749 facial identities, for the cross-dataset validation purpose. Regarding the constant coefficients for the number of blocks in Figure 2, we set $L = 4$, $M = 3$, and $N = 5$. We employed ArcFace [6] and FaceNet [29] as the assisting face recognition tools during training for the DLW computation and set $\alpha = 3$, $\beta_{\text{init}} = 0.5$, $\beta^* = 2$, $\gamma = 0.1$, $k = 30$, and $\epsilon_d = \epsilon_w = 1e-6$. For the coefficients in Eqn.(13), we set $\lambda_{\text{id}} = 0.08$, $\lambda_{\text{MSE}} = 1.8$, $\lambda_{\text{LPIPS}} = 1.2$, and $\lambda_D = 0.1$ for total loss. During training, the learning rates for NullSwap and the discriminator are $5e-4$ and $1e-4$, respectively, for 60 training epochs. Experiments are conducted on images resized to 256×256. Our model is implemented using PyTorch on 8 Tesla A100 GPUs with a batch size of 256.

4.2. Evaluation Results

We first evaluated the visual quality after embedding perturbations and discussed the performance numerically and visually. Then, we checked the identity cloaking performance of the embedded perturbation compared to state-of-the-art proactive perturbation algorithms by adopting their published best model weights. After that, ablation studies and discussions are conducted.

4.2.1. Visual Quality.

As reported in Table 1, we computed the average peak signal-to-noise ratio (PSNR), structural similarity index

Models	Clean		Initiative [16]		Anti-Forgery [32]		CMUA [15]		DF-RAP [26]		NullSwap (Ours)	
Type			Grey-Box		Grey-Box		White-Box		Grey-Box		Black-Box	
	Acc@5	Acc@1	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓
ArcFace [6]	0.980	0.976	0.974	0.968	0.977	0.975	0.979	0.976	0.977	0.974	**0.771**	**0.628**
FaceNet [29]	0.940	0.918	0.952	0.925	0.921	0.862	0.920	0.865	0.955	0.920	**0.740**	**0.590**
VGGFace [24]	0.910	0.853	0.917	0.856	0.919	0.858	0.925	0.864	0.909	0.847	**0.674**	**0.529**
SFace [2]	0.804	0.791	0.794	0.720	0.800	0.732	0.799	0.720	0.779	0.680	**0.658**	**0.513**
Average	0.909	0.885	0.909	0.867	0.904	0.857	0.906	0.856	0.905	0.855	**0.711**	**0.565**

Table 2. Top-5 and top-1 accuracies regarding the identity matching performance after applying perturbation algorithms. Lower values refer to better perturbation performance. Best performance marked in **bold**.

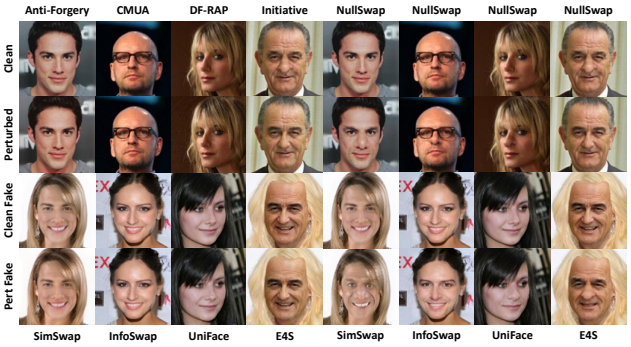


Figure 3. Visualization of face swapping performance regarding different perturbation algorithms against different generative models.

measure (SSIM), and the learned perceptual image patch similarity (LPIPS) regarding the clean and perturbed images. Specifically, considering that Initiative [16], Anti-Forgery [32], CMUA [15], and DF-RAP [26] all insert perturbations by direct element-wise addition, unavoidable visual artifacts can be perceptible to both human eyes and deep neural networks. As a result, our approach achieves state-of-the-art visual performance, with PSNR above 40 while the rest are below, SSIM above 0.98 while the rest hover around 0.95, and LPIPS below 0.005 while the rest are above 0.020.

We also visualized some samples in Figure 3 such that the first and second rows display the clean and perturbed images via different proactive perturbation approaches. It can be observed that, although achieving a PSNR above 38 and an SSIM around 0.95 guarantees acceptable visual qualities for the perturbed images, those with worse LPIPS values demonstrate artifacts in terms of illumination and blurring.

4.2.2. Identity Cloaking Performance.

Identity Cloaking on Source Faces. After deriving the perturbed image $I_s^{'}$, we tested whether $I_s^{'}$ still remains within the same identity cluster or is successfully cloaked. To do so, a face recognition task evaluation is executed on each perturbed image regarding the identity label provided by CelebA-HQ. In particular, we computed the top-5 and top-1 accuracies for the cosine similarities between identity embeddings and that of each $I_s^{'}$. To prove the generalizability of our proposed approach, besides including ArcFace [6] and FaceNet [29] that participate during the training phase, we also adopted the well-trained VGGFace [24] and SFace [2] as face recognition tools for NullSwap to fool during the testing phase.

As listed in Table 2, although values vary depending on the optimal performance of different face recognition tools, the top-5 and top-1 accuracies on clean testing images generally maintain the highest performance as expected. On the other hand, while the state-of-the-art comparative proactive perturbation models have caused accuracy degradations in most entries, the differences from the clean setting are mostly below 0.1 in absolute values except DF-RAP when facing SFace. Moreover, each of the comparative models even unexpectedly encounters greater top-5 and top-1 accuracies than that in the clean setting. This might be caused by the cases where perturbations fail to achieve the desired goal but somehow advance the feature extraction efficacy upon insertion. As a result, our proposed NullSwap achieves the average top-5 and top-1 face recognition accuracies of 0.711 and 0.565, respectively, promisingly outperforming state-of-the-art proactive perturbation algorithms such that we fooled each face recognition tool the best in the competition.

Identity Cloaking on Face Swapping. Moreover, considering the essential goal of NullSwap is to cloak the identity information and prevent Deepfake face swapping from generating images with desired identities, we tested the model performance by introducing various state-of-the-art popular face swapping algorithms, namely, SimSwap [4], InfoSwap [9], UniFace [41], E4S [23], and DiffSwap [46]. Note that the face swapping models only appear in the testing phase in our black-box setting. In this experiment, we adopted the two face recognition tools, ArcFace [6] and VGGFace [24], proved to have the best performance and stability[2], as evaluators. Specifically, to evaluate the strength of perturbations in affecting the face swapping re-

[2]https://github.com/serengil/deepface

Models	Initiative [16]		Anti-Forgery [32]		CMUA [15]		DF-RAP [26]		NullSwap (Ours)	
	ArcFace↓	VGGFace↓	ArcFace↓	VGGFace↓	ArcFace↓	VGGFace↓	ArcFace↓	VGGFace↓	ArcFace↓	VGGFace↓
SimSwap [4]	0.928	0.897	0.924	0.892	0.921	0.930	0.468	0.431	**0.217**	**0.240**
InfoSwap [9]	0.941	0.919	0.903	0.922	0.947	0.888	0.913	0.849	**0.375**	**0.359**
UniFace [41]	0.987	0.960	0.987	0.968	0.983	0.965	0.947	0.891	**0.369**	**0.329**
E4S [23]	0.925	0.893	0.924	0.901	0.920	0.900	0.880	0.855	**0.398**	**0.368**
DiffSwap [46]	0.660	0.657	0.663	0.655	0.658	0.648	0.636	0.631	**0.310**	**0.352**
Average	0.888	0.865	0.880	0.868	0.886	0.866	0.769	0.731	**0.334**	**0.330**

Table 3. Identity similarity evaluation of face swapping results on perturbed images compared to that on the clean ones using various Deepfake models via perturbation algorithms on CelebA-HQ. Best performance marked in **bold**.

Models	Clean		Initiative [16]		CMUA [15]		DF-RAP [26]		NullSwap (Ours)	
	Acc@5	Acc@1	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓	Acc@5↓	Acc@1↓
ArcFace [6]	0.923	0.910	0.864	0.801	0.925	0.903	0.891	0.841	**0.627**	**0.491**
FaceNet [29]	0.822	0.708	0.726	0.555	0.822	0.702	0.773	0.621	**0.324**	**0.210**
VGGFace [24]	0.704	0.550	0.588	0.429	0.701	0.555	0.648	0.494	**0.294**	**0.186**
SFace [2]	0.553	0.445	0.485	0.369	0.565	0.455	0.542	0.437	**0.348**	**0.225**
Average	0.751	0.653	0.666	0.539	0.753	0.654	0.714	0.598	**0.398**	**0.278**
PSNR↑	–		34.1381		37.8710		30.6551		**42.8132**	
SSIM↑	–		0.8380		0.9392		0.8824		**0.9882**	
LPIPS↓	–		0.3102		0.0517		0.1422		**0.0049**	

Table 4. Top-5 and top-1 accuracies regarding the identity matching performance after applying perturbation algorithms on LFW. Lower values refer to better perturbation performance. Best performance marked in **bold**. Anti-Forgery [32] is omitted since LFW does not contain the attribute information it requires.

sults, we first generated face swapping results via each generative model using clean images. Then, we repeated the same pipeline along with the identical source-target pairs using the perturbed images derived by each proactive perturbation model. For face swapping results based on perturbed faces, we compute the cosine similarity with respect to the corresponding clean face swapping results based on their identity embeddings.

As reported in Table 3, Initiative [16], Anti-Forgery [32] and CMUA [15] are observed to derive face swapping results that have cosine similarities all around 0.9, except DiffSwap [46], regarding the clean ones, regardless of specific face swapping algorithms. This implies that, although demonstrating reasonable defensive ability against easier tasks such as attribute editing, as reported in the original work, the inserted perturbations generally fail to distort or nullify Deepfake face swapping. On the other hand, DF-RAP [26], being evaluated against SimSwap in the original work, appears to be effective in nullifying SimSwap [4], with low cosine similarities of 0.468 and 0.431 for ArcFace and VGGFace as expected. However, the perturbations are observed to lack generalizability toward other face swapping models, leading to similarities around or above 0.850. DiffSwap [46], a diffusion-based face swapping method, although executes unstable generation results due to its original performance, is still highly nullified by our approach, deriving identity similarities as low as 0.310 and 0.352 for ArcFace and VGGFace, respectively. As a result, our proposed method consistently nullifies different state-of-the-art face swapping models by cloaking the identity information, achieving average similarities of 0.334 and 0.330 for ArcFace and VGGFace, respectively, with individual similarities all below 0.4.

This is consistent with the visualized results in Figure 3. Particularly, the perturbed fake faces are highly identical to the clean fake faces in the columns of Anti-Forgery, CMUA, and DF-RAP, and no visible distortion can be observed. As for Initiative, only lighting artifacts can be observed. Contrarily, in the first two columns of our proposed NullSwap, obvious visual differences are observed between the clean and perturbed face swapping results. For the rest two columns, detailed but critical identity-specific differences are caused by the identity cloak on the nose, mouth, and cheek areas.

Cross-Dataset Validation. To further validate the generalizability of our proposed NullSwap on unseen data, we conducted cross-dataset validations on the Labeled Faces in the Wild (LFW) [14] dataset. Experiments are conducted following the same pipeline as Table 2 except Anti-Forgery is omitted due to unprovided attribute information in LFW. As Table 4 shows, while CMUA [15] generally maintains similar top-5 and top-1 identity recognition accuracies as the

Models	Session 1	Session 2	Session 3	DLW (Ours)
ArcFace↓	0.653	0.843	0.846	0.628
FaceNet↓	0.758	0.546	0.613	0.590
VGGFace↓	0.680	0.504	0.600	0.529
SFace↓	0.576	0.495	0.525	0.513
Average	0.667	0.597	0.646	0.565
PSNR↑	42.1461	40.3650	40.8079	41.3097
SSIM↑	0.9872	0.9814	0.9747	0.9864
LPIPS↓	0.0042	0.0075	0.0063	0.0049

Table 5. Ablation study on the strategies regarding identity losses. The NullSwap model is trained with Session 1 (ArcFace loss), Session 2 (FaceNet loss), Session 3 (average of ArcFace and FaceNet losses), and the proposed Dynamic Loss Weighting (DLW).

clean ones for all four face recognition tools, Initiative [16] and DF-RAP [26] achieve the defensive goal at certain levels but fail to preserve reasonable visual qualities. On the other hand, our proposed identity cloaking strategy leads to the largest degradations with 0.398 and 0.278 on average for top-5 and top-1 accuracies, while consistently maintaining promising image visual quality. In conclusion, NullSwap demonstrates outstanding generalizability on unseen datasets such as LFW.

4.2.3. Ablation Study

In this section, we validated the performance of our proposed Dynamic Loss Weighting (DLW) strategy regarding the case when introducing multiple face recognition tools to help ensure generalizability. We conducted four training sessions in different settings with respect to the participating face recognition tools in this study. Specifically, Session 1 adopts ArcFace [6] solely to provide cosine similarity to identity loss, and similarly, Session 2 adopts FaceNet [29] solely. In Session 3, both ArcFace and FaceNet are adopted, and identity loss is computed by averaging cosine similarities. Lastly, we kept our optimal DLW in Session 4.

Since the only difference happens in the identity loss, the visual quality of all sessions does not heavily fluctuate. Therefore, we focused on analyzing the identity cloaking performance of different sessions. Following the same evaluation process as in Table 2, the top-1 face recognition accuracy of each training session is exhibited in Table 5. Session 1, although successfully decreases the accuracy on ArcFace, comes up with the highest accuracies on FaceNet, VGGFace, and SFace, leading to the largest average accuracy of 0.667. On the other hand, despite maintaining high accuracy for ArcFace, Session 2 trained with FaceNet is able to drag the accuracies of the rest down to lower levels. This might be due to the feature overlapping of the way they derive identity embeddings. However, training by simply averaging the loss for ArcFace and FaceNet in Session 3 fails to achieve better performance. Contrarily, our proposed Dynamic Loss Weighting performs consistently for all face recognition tools. This proves the fact that identity losses of ArcFace and FaceNet suffer fluctuations between batches during the training session, which critically affects the achievement of the objectives. Furthermore, DLW considers history losses, which not only stabilizes the fluctuations among batches but also adaptively adjusts weight updating regarding loss values that are stuck at certain points. As a result, although not achieving the optimal results for all entries, our optimal session with DLW achieves the best average performance on identity cloaking at 0.565.

4.2.4. Discussion

Deepfake threats have happened in recent years in cases such as false scandals and economic fraud [38] relying on the success of reconstructing the source face identity. While non-professionals usually have tolerance for the visual quality upon mere glances but tend to believe the truthfulness once recognize the facial identities, nullifying Deepfake face swapping and making the generated images with undesired identities can effectively resolve the threats. In this study, despite the original accuracies of face recognition tools can vary and the performance of face swapping algorithms may be unstable, our proposed method consistently decreases the similarities between the cloaked identity information and the clean ones while maintaining acceptable visual quality, as demonstrated in Tables 1 to 4. Besides, due to the diversity in existing and future Deepfake face swapping algorithms, the white- and grey-box settings with generative models participating during training can progressively become computation consuming while still lacking generalizability. As proved in Sections 4.2.2 and 4.2.3, introducing DLW for multiple identity recognition tools consistently achieves promising performance across seen and unseen face recognition tools and face swapping models.

5. Conclusion

In this study, we analyze the essence of the Deepfake face swapping attack and claim the necessity of protecting the source images that provide facial identities, the real victims, rather than the target images that are visually edited. We propose an identity cloaking proactive approach, NullSwap, to adaptively embed identity-guided perturbations into source images for nullifying face swapping such that synthetic faces are generated with undesired facial identities. In addition to preserving reasonable image visual quality, our approach demonstrates the ability to fool various identity extractors of different face swapping algorithms in cross-dataset and cross-model scenarios. Besides outperforming state-of-the-art proactive perturbations, the novel claim of protecting the source identities rather than target images provides innovative insights, and the black-box scenario without generative models participating during training deserves further deep analyses in future studies.

Acknowledgments

This work was supported by Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences).

References

[1] Shivangi Aneja, Lev Markhasin, and Matthias Nießner. Tafim: Targeted adversarial attacks against facial image manipulations. In *Computer Vision – ECCV 2022*, pages 58–75, 2022. 2

[2] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics*, pages 1–11, 2022. 6, 7

[3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 1

[4] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *The 28th ACM International Conference on Multimedia*, page 2003–2011, 2020. 2, 3, 6, 7

[5] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection, 2024. 1

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 2, 3, 5, 6, 7, 8

[7] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023. 2

[8] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 1

[9] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3403–3412, 2021. 2, 6, 7

[10] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[11] Ziwen He, Wei Wang, Weinan Guan, Jing Dong, and Tieniu Tan. Defeating deepfakes via adversarial visual reconstruction. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2464–2472, 2022. 3

[12] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, 2002. 1

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 4

[14] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to align from scratch. In *Conference on Neural Information Processing Systems*, 2012. 5, 7

[15] Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):989–997, 2022. 2, 5, 6, 7

[16] Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1, 2, 5, 6, 7, 8

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision*, pages 1510–1519, 2017. 2

[18] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision*, 130(7):1678–1734, 2022. 1

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014. 2

[21] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, 2017. 2

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 5

[23] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8578–8587, 2023. 2, 6, 7

[24] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 3, 6, 7

[25] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework, 2021. 2

[26] Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network

scenarios. *IEEE Transactions on Information Forensics and Security*, 2024. 3, 5, 6, 7, 8

[27] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision – ECCV 2020 Workshop*, page 236–251, 2020. 1, 2

[28] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1, 3

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 3, 5, 6, 7, 8

[30] Zhimin Sun, Shen Chen, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Rethinking open-world deepfake attribution with multi-perspective sensory learning. *International Journal of Computer Vision*, 133:628–651, 2024. 1

[31] Long Tang, Dengpan Ye, Zhenhao Lu, Yunming Zhang, and Chuanxi Chen. Feature extraction matters more: An effective and efficient universal deepfake disruptor. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(2), 2024. 2

[32] Run Wang, Ziheng Huang, Zhikai Chen, Li Liu, Jing Chen, and Lina Wang. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 761–767, 2022. 1, 2, 5, 6, 7

[33] Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14548–14556, 2023. 1

[34] Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. Deep convolutional pooling transformer for deepfake detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(6), 2023. 1

[35] Tianyi Wang, Mengxiao Huang, Harry Cheng, Bin Ma, and Yinglong Wang. Robust identity perceptual watermark against deepfake face swapping, 2023. 1

[36] Tianyi Wang, Mengxiao Huang, Harry Cheng, Xiao Zhang, and Zhiqi Shen. Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks. *The 32nd ACM International Conference on Multimedia*, 2024. 1

[37] Tianyi Wang, Zian Li, Ruixia Liu, Yinglong Wang, and Liqiang Nie. An efficient attribute-preserving framework for face swapping. *IEEE Transactions on Multimedia*, 26:6554–6565, 2024. 2

[38] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 2024. 1, 8

[39] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. Deepfake disrupter: The detector of deepfake is my friend. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14900–14909, 2022. 1, 2

[40] Xiaoshuai Wu, Xin Liao, and Bo Ou. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 1

[41] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *European Conference on Computer Vision*, pages 54–71, 2022. 2, 3, 6, 7

[42] Chaofei Yang, Leah Ding, Yiran Chen, and Hai Li. Defending against gan-based deepfake attacks via transformation-aware adversarial faces. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. 2

[43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[44] Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Xin Liao, Ziyi Liu, Chuanxi Chen, and Jiacheng Deng. Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping. *IEEE Transactions on Information Forensics and Security*, 19:4628–4641, 2024. 3

[45] Yaning Zhang, Zitong Yu, Tianyi Wang, Xiaobin Huang, Linlin Shen, Zan Gao, and Jianfeng Ren. Genface: A large-scale fine-grained face forgery benchmark and cross appearance-edge learning. *IEEE Transactions on Information Forensics and Security*, pages 8559–8572, 2024. 1

[46] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023. 6, 7

[47] Yao Zhu, Yuefeng Chen, Xiaodan Li, Rong Zhang, Xiang Tian, Bolun Zheng, and Yaowu Chen. Information-containing adversarial perturbation for combating facial manipulation systems. *IEEE Transactions on Information Forensics and Security*, 18:2046–2059, 2023. 2